xr9caulzy

August 22, 2024

1 Model Selection Homework

Aug 22, 2024

1.1 Bài toán

- Dự đoán giá nhà
- Dữ liệu gồm thông tin căn nhà và giá
- So sánh các mô hình: Hồi quy Lasso, Hồi quy Ridge, Hồi quy tuyến tính, Rừng ngẫu nhiên và KNN

1.2 Import thư viện cần thiết

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from tqdm import tqdm

from sklearn.model_selection import learning_curve
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import Pipeline

from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.ensemble import RandomForestRegressor
%matplotlib inline
```

[ ]:
```
import sklearn
sklearn.metrics.get_scorer_names()
```

1.3 Đọc dữ liệu

[ ]:
```
%cd /content/drive/MyDrive/Code_VinBigData_2024/Model_selection
```

[ ]:
```
dataset = pd.read_csv('kc_house_data.csv')
dataset.head()
```

[ ]:
```
dataset.info()
```

[ ]:
```
dataset.describe()
```

1.4 Tiền xử lý dữ liệu

[ ]:
```
###Code here
```

1.5 Chia dữ liệu Train - Test

Tỷ lệ Train - Test = 8 - 2

[ ]:
```
###Code here
```

1.6 Các hàm quan trọng

[ ]:
```
def cross_validation(estimator):
    _, train_scores, test_scores = learning_curve(estimator,
                                                  X_train, Y_train,
                                                  cv=10,
                                                  n_jobs=-1,
                                                  train_sizes=[1.0, ],
                                                 ␣
 ↪scoring='neg_mean_absolute_error')
    test_scores = test_scores[0]
    mean, std = test_scores.mean(), test_scores.std()
    return mean, std

def plot(title, xlabel, X, Y, error, ylabel = "mean_squared_error"):
    plt.xlabel(xlabel)
    plt.title(title)
    plt.grid()
    plt.ylabel(ylabel)

    plt.errorbar(X, Y, error, linestyle='None', marker='o')
```

1.7 Lựa chọn tham số cho các mô hình

[ ]:
```
###Code here
```

1.8 So sánh các mô hình

[ ]:
```
###Code here
```

[ ]:
```
# Kết quả dự đoán trên tập test
print(f'RF: {mean_absolute_error(Y_test, rf.predict(X_test))}')
print(f'KNN: {mean_absolute_error(Y_test, knn.predict(X_test))}')
print(f'Linear Regression: {mean_absolute_error(Y_test, lrg.predict(X_test))}')
print(f'Ridge: {mean_absolute_error(Y_test, ridge.predict(X_test))}')
print(f'Lasoo: {mean_absolute_error(Y_test, lasso.predict(X_test))}')
```